LEAST SQUARES LINEAR REGRESSION ANALYSIS

In this unit we investigate fitting a straight line to measured (x, y) data pairs. The equation we want to fit is of the form:

$$y = ax + b$$

where a is the slope and b is the intercept. If we only have two data pairs we can fit a unique line to them. However, when we have more data pairs we could fit a large number of different lines through the scatter of data points. The problem therefore arises: how do we choose the "best" straight line?

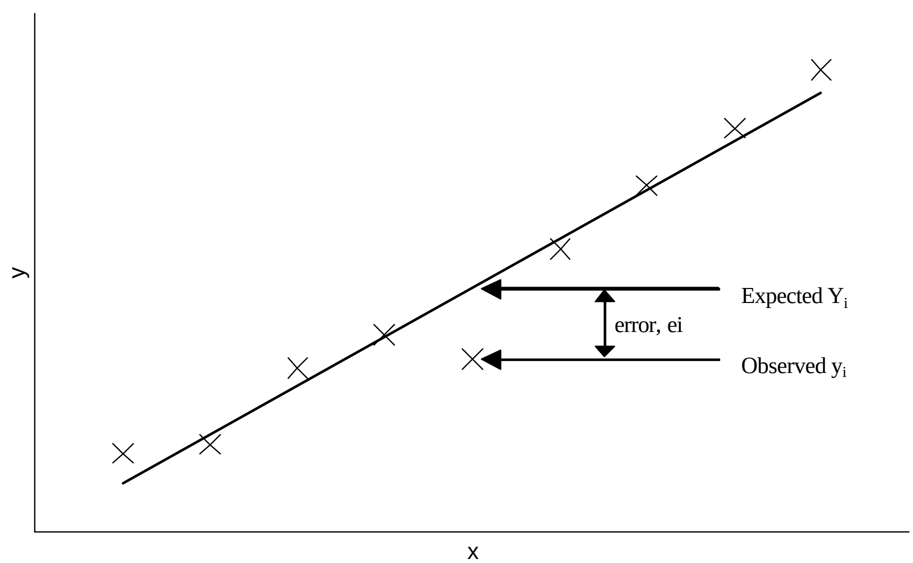
There are several different methods. The one we develop here is the least squares error approach. This is the one most commonly used in engineering and is readily available in Excel, MathCAD and your calculator.

Let us assume we have n data pairs (x_i, y_i) with i = 1 .. n. **We make the assumption that all *x* values have no error. The only error is associated with the *y* values.** If we knew the slope and intercept of our best-fit line, we could predict the expect y-value for any given x-value. We choose upper case Y to represent the expected values and lower case y to represent the measure values:

$$Y_i = ax_i + b$$

There will be an error, e_i, (difference) between the expected values and observed values:

$$\begin{aligned} e_i &= Y_i - y_i \\ &= ax_i + b - y_i \end{aligned}$$



We will try to minimize the sum of the squares of the errors, *E*.

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - y_i\right)^2 = \sum_{i=1}^{n} \left(ax_i + b - y_i\right)^2$$

Remember that at this instant we do not yet know the slope or intercept, so we cannot do any numerical calculations with our data.

Let's press on. We wish to minimize the error *E* by choosing the 'best' slope and intercept. We do this by differentiating the error function wrt slope and intercept (separately) and setting the result to zero to minimize the total error:

$$\text{wrt slope:} \quad \frac{\partial E}{\partial a} = 0 = \sum_{i=1}^{n} 2\left(ax_i + b - y_i\right)x_i$$

$$\text{wrt intercept:} \quad \frac{\partial E}{\partial b} = 0 = \sum_{i=1}^{n} 2\left(ax_i + b - y_i\right)$$

We solve these two equations simultaneously to get:

$$\text{slope:} \quad a = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\text{intercept:} \quad b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

or

$$\boxed{\begin{array}{l} \text{slope:} \quad a = \dfrac{\sum_{i=1}^{n} x_i y_i - n^2 \bar{x}\,\bar{y}}{n\sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2} \\[2em] \text{intercept:} \quad b = \dfrac{n\sum_{i=1}^{n} x_i^2 \bar{y} - n^2 \bar{x}\,\bar{y}}{n\sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2} \end{array}}$$

The resulting line $y = ax + b$ is called the *least-squares best fit line* or just the "*best-fit line*", to the data represented by the data set (x_i, y_i).

One estimate of how well the best-fit line represents the data is the *standard error of the estimate.* For practical purposes we can consider this to be the standard error of the intercept. MathCAD has a function *stderr* that will calculate the quantity. Excel and your calculator do not! The quantity is calculated as:

$$S_{y,x} = \sqrt{\frac{\sum y_i^2 - b\sum y_i - a\sum x_i y_i}{n-2}}$$

You can calculate a confidence interval for the intercept using methods similar to the work we developed for the Normal distribution and parameter estimation:

$$d_{INTERCEPT} = tS_{y,x}$$

or

$$\text{intercept} = b \pm d_{INTERCEPT} = b \pm tS_{y,x}$$

where *t* is the Student's-t for the appropriate degrees of freedom (n – 2) and level of significance, α.

IMPORTANT: Never just quote the least-squares slope and intercept. ALWAYS generate a graph that shows the original data and the reconstructed best-fit line. Normally you should also include the two lines that show the best-fit line $\pm d_{INTERCEPT}$.

Standard Error of the Slope:

There is also an equation that determines the standard error of the slope. It is not available in either MathCAD or Excel, so if you need to use it, you will have to 'hard-code' it yourself. The standard error of the slope is:

$$S_{a1} = S_{y,x}\sqrt{\frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2}}$$

The confidence interval of the slope is:

$$d_{SLOPE} = tS_{a1}$$

Again, *t* is the Student's-t for the same degrees of freedom (n – 2) and level of significance, α, used to calculate the standard error of the intercept, $S_{y,x}$.

PROCEDURE FOR DETERMINING A BEST-FIT STRAIGHT LINE USING

LINEAR
CORRELATION AND REGRESSION

1. Plot the raw data
2. Make a subjective decision about the data and any trends

3. Calculate correlation coefficient r_{xy} and compare it to the values in the table. Make a statistical decision

4. Calculate slope & intercept of the best-fit line
5. Calculate standard error of estimate
6. Calculate confidence interval of the intercept

7. Plot raw data and overlay the best-fit line with the ±confidence intervals

Remember that most data points MUST fall inside the CI
Expect a few points to fall outside the CI

FINAL DECISION BASED ON
SUBJECTIVE – LOOK LIKE A STRAIGHT LINE?
STATISTICAL – CORRELATION COEFFICIENT

EXAMPLE: This example was developed in MathCAD. It demonstrates the steps necessary to do a linear, least-squares regression.

Example on Linear Regression using MathCAD

D :=

	0	1
0	-4.98	-11.91
1	-2.1	-17.17
2	3.78	21.03
3	0.25	6.07
4	7.34	4.78
5	-2.39	-3.91
6	5.66	43.91
7	-0.44	-0.42
8	-3.63	-17.12
9	-2.79	-20.9

This matrix of raw data was read from a file using an import table

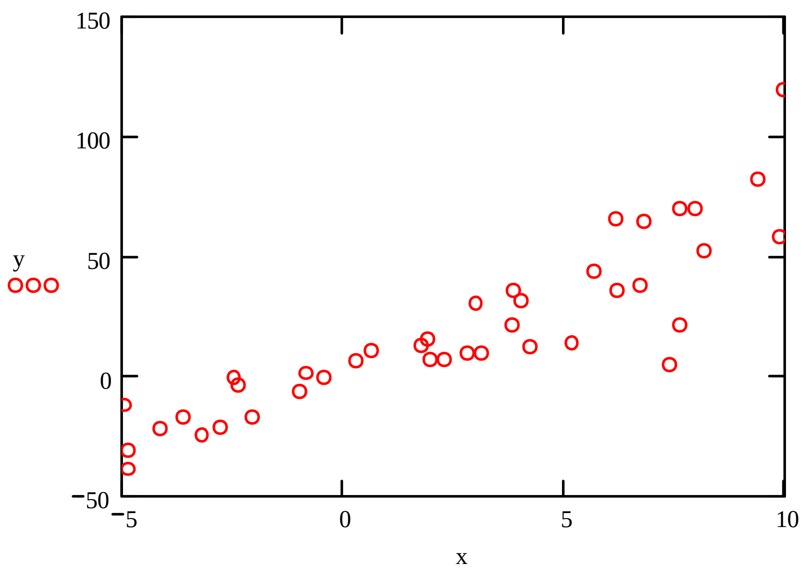
$n := \mathrm{rows}(D) \qquad n = 40$

$x := \mathrm{submatrix}(D, 0, n-1, 0, 0)$

$y := \mathrm{submatrix}(D, 0, n-1, 1, 1)$

Extract the x- and y- data into separate vectors

Plot the raw data to see what it looks like



There might be a linear trend, but there is also some scatter to the data.

Calculate the slope and intercept of the best-fit line

$\mathrm{slp} := \mathrm{slope}(x, y) \qquad \mathrm{slp} = 6.6863$

$\mathrm{interc} := \mathrm{intercept}(x, y) \qquad \mathrm{interc} = 2.1828$

Calculate the correlation coefficient and standard error

$\mathrm{corr}(x, y) = 0.8838$

$\mathrm{stdE} := \mathrm{stderr}(x, y) \qquad \mathrm{stdE} = 16.2416$

We comapre the correlattion coefficent with tabulated values.

Calculated = 0.8838 Tabulated (at 5% level of significance and n=40) = 0.312
Since calculated > tabulated, we conclude that there is no strong evidence that a trend does not exist

Now on to the confidence interval. Look up Student's-t for dof=n-2

$t := qt(.975, n-2)$ $\qquad t = 2.0244$ $\qquad$ in the 'qt' function, the first number is 1-(α/2)

$CI := t \cdot stdE$ $\qquad$ CI is the confidence interval $\qquad CI = 32.8794$

Set up everything for plotting

$yLS(xLS, d) := interc + slp \cdot xLS + d$ $\qquad$ The straight line equation with'extra' intercept

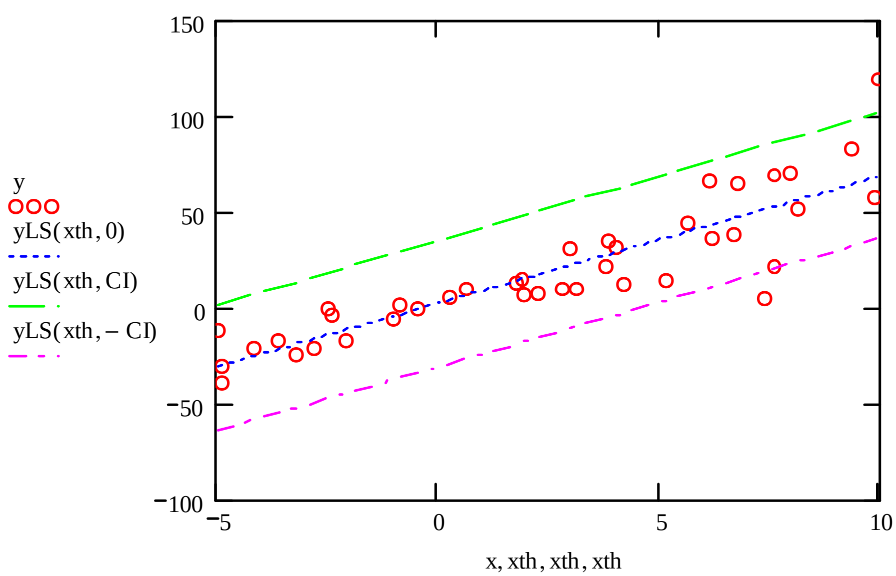
In this example, the best fit line and +/- confidence interval lines are plotted by defining their start and end points (2 points define a line!!)

$xth_0 := \min(x)$ $\qquad$ This is the smallest X value, and is the first x-point for our fitted lines

$xth_1 := \max(x)$ $\qquad$ This is the biggest (and second) x-point for our least squares lines

$$xth = \begin{pmatrix} -4.9810 \\ 9.9519 \end{pmatrix}$$

Plot the raw data, best-fit line and confidence interval



I have left this plot with ARGUMENTS showing and LEGEND hidden so you can see exactly how I generated the plot.

Conclusions for this example:

The plot of the raw data was inconclusive. There may be a linear trend, but there is also significant scatter.

The correlation coefficient is well above the tabulated values. This hints toward a linear trend

As expected, most of the raw data points fall inside the 95% confidence interval.

> QUESTION: How many data points would you *expect* to fall outside the confidence interval?
>
> What does it mean if you don't get any points (or if you get lots of points) outside the confidence interval?

The change in y-value predicted by the best-fit line changes (from smallest to largest x-values) by more than the confidence interval. Therefore, the trend predicted by the line has greater variation than the variation caused by the noise in the data.

Background knowledge and theory may help determine whether the data are expected to follow a straight line or not.

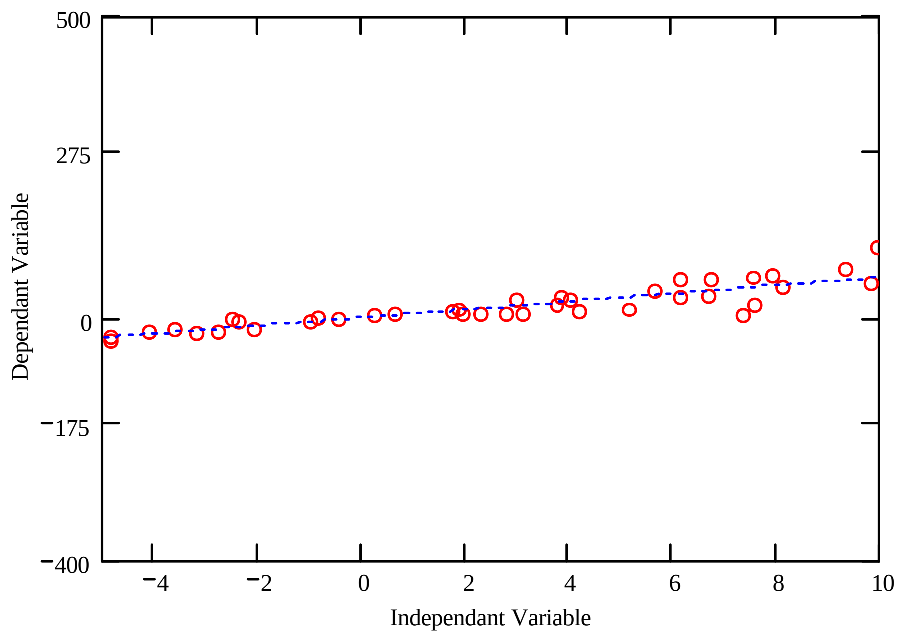
My final summary for this data set is that a straight line can be used to predict the output (y) for a given input (x) within the range $-5 < x < 10$. The 95% confidence interval on y-values is about 32.9 (in measurement units).

QUIZ TIME

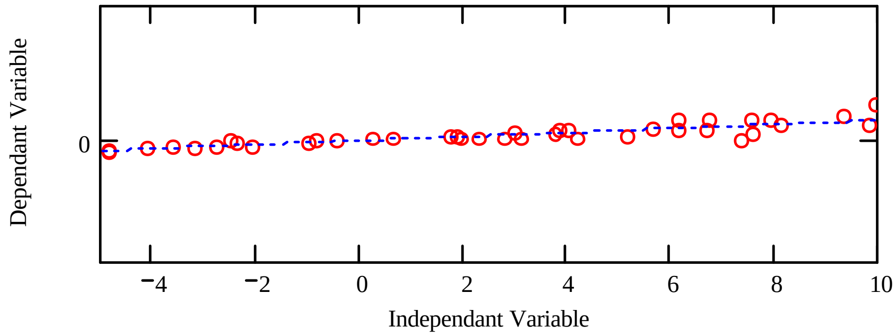
There are 5 graphs, A through E.

You will have about 5 seconds per graph. Each graph shows some raw data with the best-fit straight line going through the data. Decide if the straight line is a good representation of the data. I.e., do the data follow a straight line?

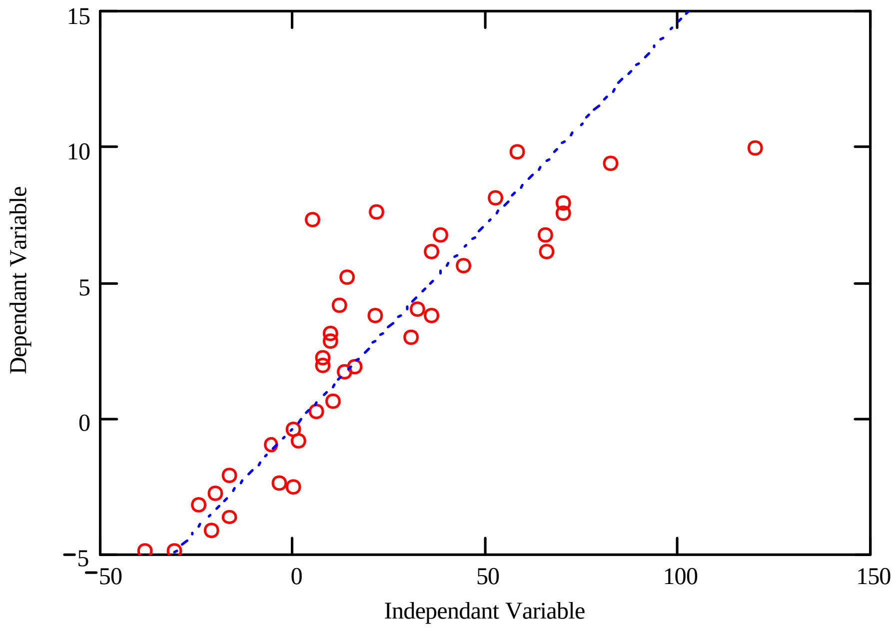
GRAPH A



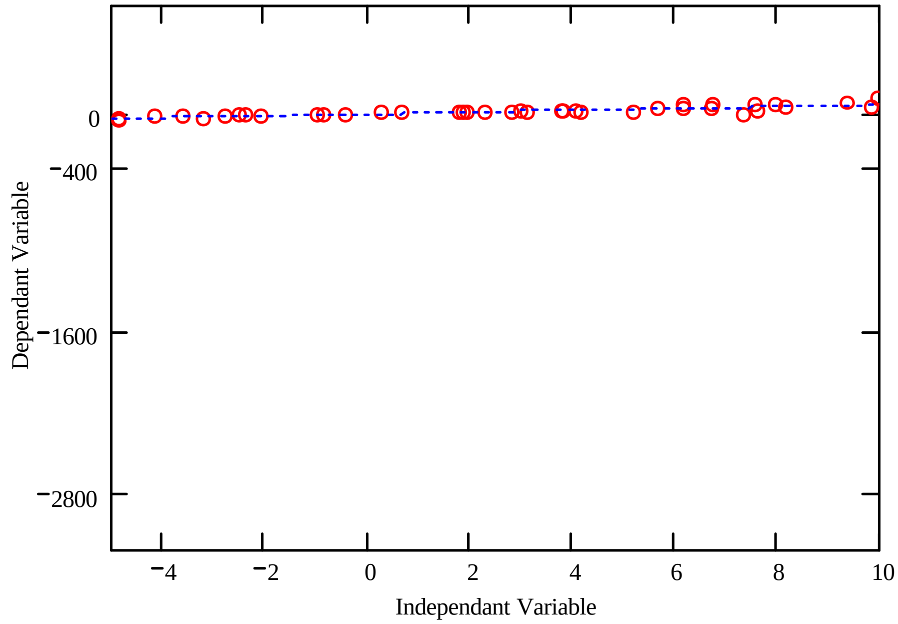
GRAPH B



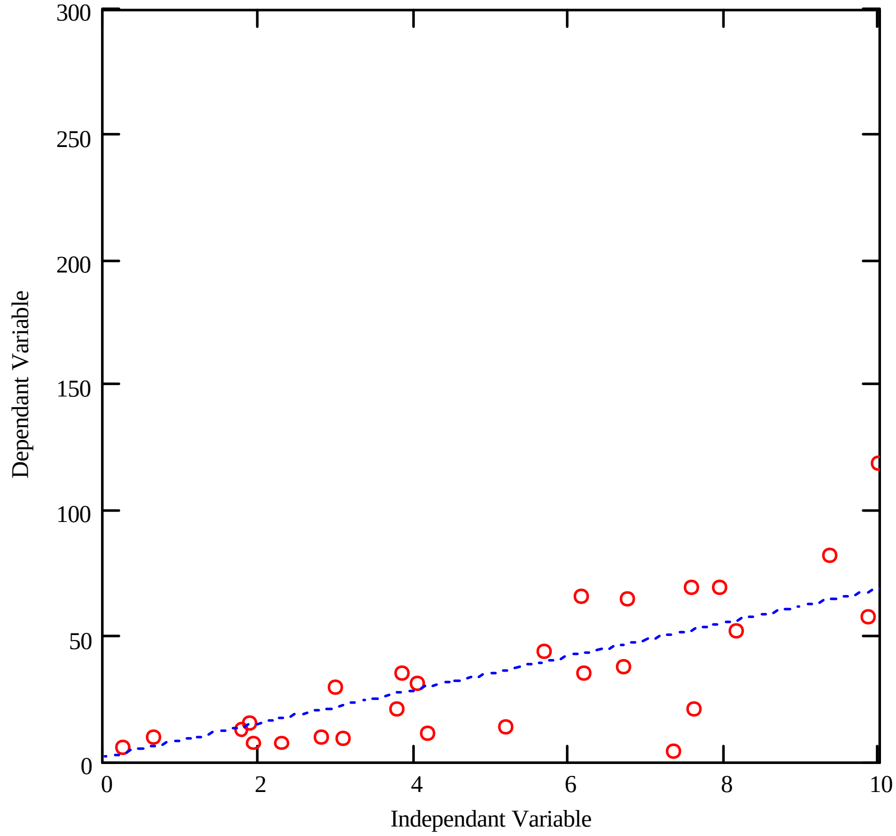
GRAPH C


15
10
5
0
−5
Dependant Variable
−50
0
50
100
150
Independant Variable

GRAPH D


0
−400
−1600
−2800
Dependant Variable
−4
−2
0
2
4
6
8
10
Independant Variable

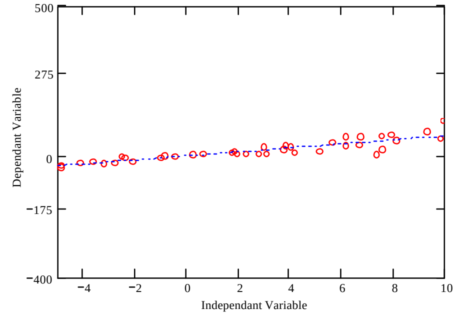
GRAPH E

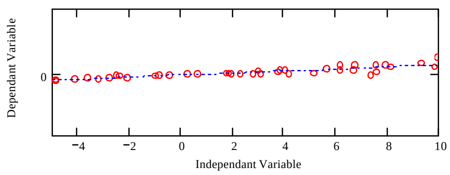


300
250
200
150
100
50
0
Dependant Variable
0
2
4
6
8
10
Independant Variable

How well did you do? The answer is that all 5 graphs showed the **same** x-y data.

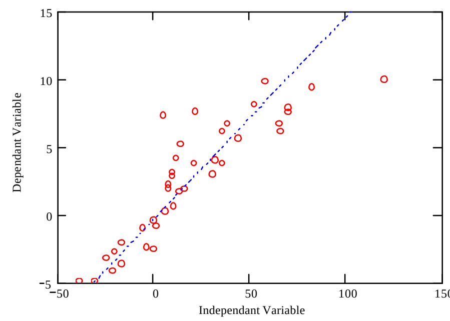
Graph A had a compressed vertical scale, minimizing the visual effect of the scatter on the data.

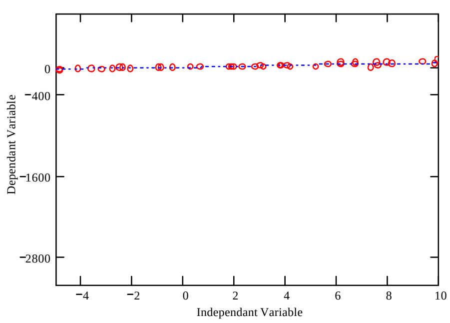




Graph B has the same axis limits as Graph A, except the physical size of the graph was reduced

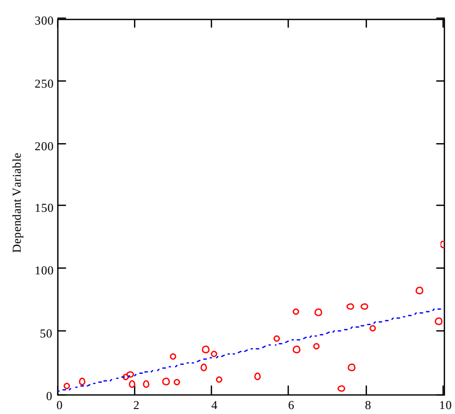
Graph C used auto scaling for both axes.





Graph D had a very large vertical axis scaling, making the scatter almost impossible to see.

Graph E flipped the x-y axes and only plotted positive values.



What lessons did you learn from this quiz?